Published in: Proceedings of the Tenth European Association for Lexicography. Copenhagen Vol. 1 - Euralex, 2002. Pp. 359 - 369

Collocational Information in the FrameNet Database

Josef Ruppenhofer, Collin F. Baker, and Charles J. Fillmore

International Computer Science Institute
1947 Center St.
Berkeley, CA 94704-1198, USA
{josef, collinb, fillmore}@icsi.berkeley.edu
website: http://framenet.icsi.berkeley.edu/~framenet

Abstract

The FrameNet lexical database yields information about collocations and multiword expressions in various ways. In some cases phrasal units have been entered from the start as lexical entries (*write down*). In other cases headword+preposition pairs can be recognized as special collocations where the preposition in question is a necessary and lexically specified marker of an argument of the headword (*fond of, hostile to*). Nominal compounds are annotated with respect to noun or (pertinative) adjective modifiers, some of which are analyzable but also entrenched (*wheel chair, fiscal year*). Nouns that name aggregates, portions, types, etc., sometimes hold lexically specified relations to their dependents (*flock of geese*). And event nouns frequently select the support verbs which permit them to enter into predications (*file an objection, enter a plea*). A subproject aims at extracting, as structured clusters of lexical items, the minimal semantically central **kernel dependency graphs** from the set of annotations. Such research will yield not only commonplace groupings (eat: dog, bone) but will also yield hitherto unnoticed collocations within such graphs (answer: you, door) where certain dependency links within them are idiomatic or otherwise lexically special, here *answer* > *door*. Collocational information can also be retrieved by various types of queries within our MySQL search tool.

Introduction

The FrameNet research project [Baker et al. 1998; Fillmore & Baker 2001] is building an online lexical resource that aims to provide, for a significant portion of the vocabulary of contemporary English, a body of semantically and syntactically annotated sentences from which reliable information can be reported on the valences or combinatorial possibilities of each item included. The project uses a descriptive model based on semantic frames [Fillmore 1977, 1982, 1985; Fillmore & Atkins 1988] and documents its observations by means of carefully annotated attestations taken from corpora, each sentence annotated in respect to a single **target word** with the phrases that are in grammatical construction with it labelled according to their grammatical relation to the target, the semantic role they serve within the frame to which the target word belongs, and its syntactic phrase type.

The FN database can serve both human and machine "users" and can function both as a dictionary and as a thesaurus. As a dictionary, each **lexical unit** (lemma in a given sense) is provided with (1) the name of the frame it belongs to and access to a description of the frame, (2) a definition (either original or from the Concise Oxford Dictionary, courtesy of Oxford University Press), (3) a valence description which summarizes the attested combinatorial possibilities in respect to both semantic roles and the syntactic form and function of the phrases that instantiate those roles, and (4) access to annotated examples illustrating each syntactic pattern found in the corpus and the kinds of semantic information they contain. The semantic role annotation is done manually by persons trained in frame

semantic theory; the syntactic information is added automatically, and the full valence descriptions are produced automatically.

It is possible to consider the database as a thesaurus by noting that lemmas are linked to the semantic frames in which they participate, and frames, in turn, are linked both to the full set of words which instantiate them and to related frames. Frame-to-frame relations include (1) **composition**, by which a complex frame is decomposable into subframes, often in a structured procedural sequence (thus, the Arraignment frame is treated as a subframe of the Criminal Justice frame), and (2) **inheritance**, by which a single frame can be seen as an elaboration of one or more other frames, with bindings between the inherited semantic roles (as when *criticize* (in the Judgement_Communication frame) can be seen as inheriting from both the Judgement and Communication frames, requiring a binding between the Speaker of the Communication frame and the Judge of the Judgement.

The FrameNet Database

The FN data are stored in a MySQL database [Fillmore et al. 2001] which is basically divided into two halves, one representing the frames, the frame elements, the lemmas connected with them and the relations among them (shown in Fig. 1), and the other (not shown) representing the corpus sentences and the labels attached to them, marking phrases as instantiations of given frame elements, phrase types and grammatical functions, etc. This division corresponds to the two main software tools used in FN work, the Frame Editor and the Annotation tool, both of which will be demonstrated in the FrameNet demonstration.

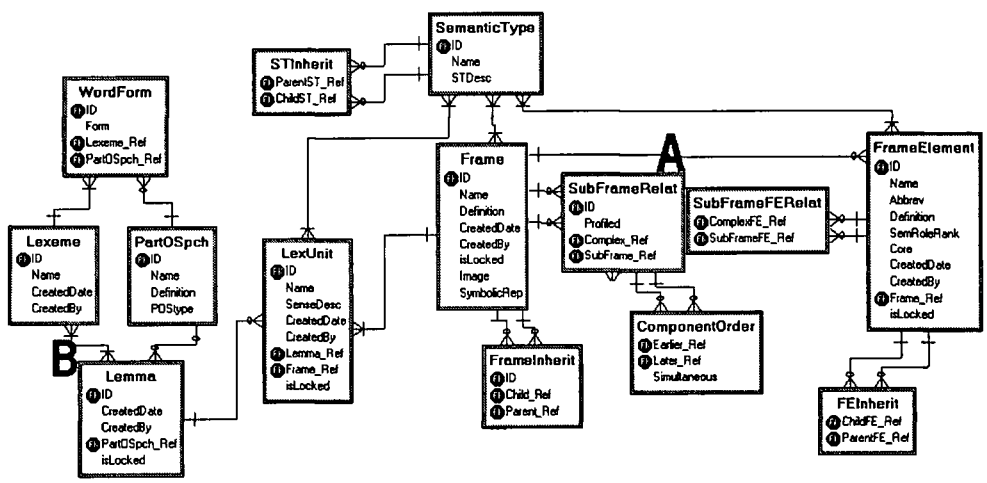

Figure 1. The Structure of the FrameNet Database (partial)

The structure of the database is designed to represent frame structures and frame-theoretic relations within the lexicon as directly as possible, although in a few places deviations from this principle have been allowed for the sake of computational efficiency. Each of the fundamental units of analysis is represented by a table. Thus, there is a table for Frames

including the name and definition, one for frame elements, and a relation between them (the line marked "A" in Fig. 1) such that each FE is associated with exactly one frame. Lexical units (LUs) are represented as a table linking lemmas and frames, i.e. an LU is a Saussurean sign linking form and meaning; there is also a field for a description of the sense. Lemmas, in turn are composed of one or more lexemes, and lexemes have one or more word forms. For example, the lexeme *grill* (with word forms *grilled*, *grilling*, etc.) is the only lexeme of a lemma which is associated with two quite different frames, **Questioning** and **Apply_heat**.

Of particular interest to the present discussion is the relation between lemmas and lexemes, which is many to many (the line marked "B" in Fig. 1), meaning that a lemma can be comprised of more than one lexeme (multiword expressions, MWEs), and a lexeme can be associated with more than one lemma. For example, the lemmas *write up, write down* and *write in* are all associated with the Writing frame on the one hand, and all contain the lexeme *write*, sharing its word forms *writing*, *wrote*, etc. on the other.

There are three procedures by which data are made part of the database: (1) the annotation process, through which, for a specifically targeted lexical unit, exemplary sentence constituents are tagged according to their semantic and syntactic relation to the target; (2) descriptions of frames prepared by the frame analyst, and (3) relations among frames prepared by the lexicon analyst.

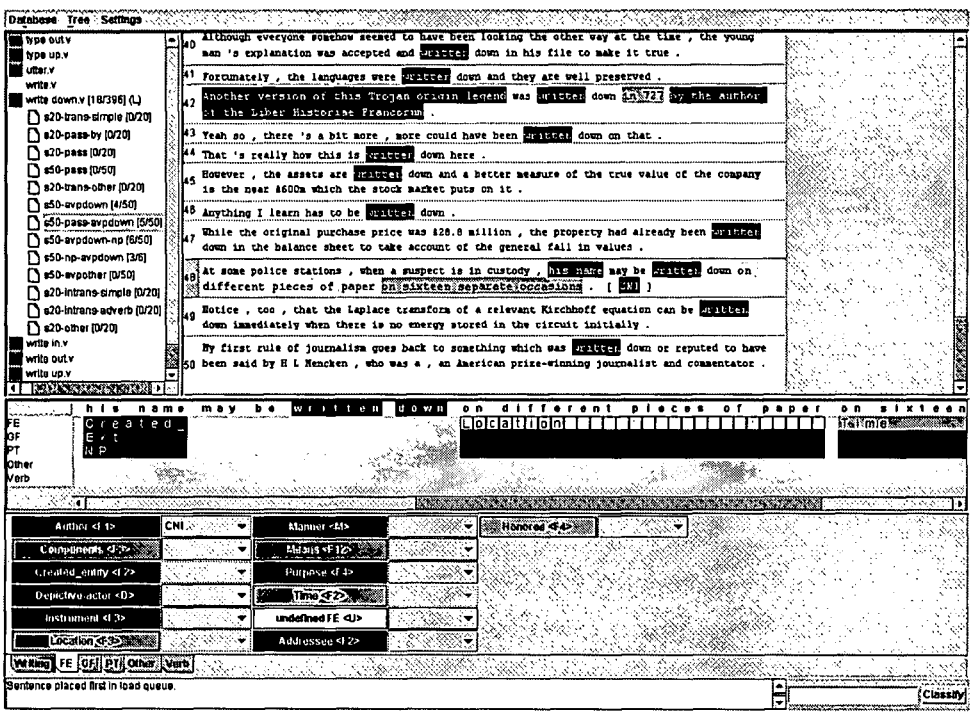

Figure 2. Screenshot of the Annotation Tool

The Annotator, shown in Figure 2, consists of three frames[1]. The top left one contains a listing of the names of individual subcorpora within which annotation is carried out, derived from searches for the target word (*write*) in certain predefined syntactic contexts. The top right frame lists the sentences of the currently selected subcorpus. The lower frame is the place where individual sentences are annotated by applying (using the mouse or the keyboard) labels to the annotation layers of the constituents instantiating particular semantic roles. The most important annotation layers are the top three shown here, which represent the Frame Element (i.e. frame-specific semantic role), the Grammatical Function, and the Phrase Type of the constituent being labeled.

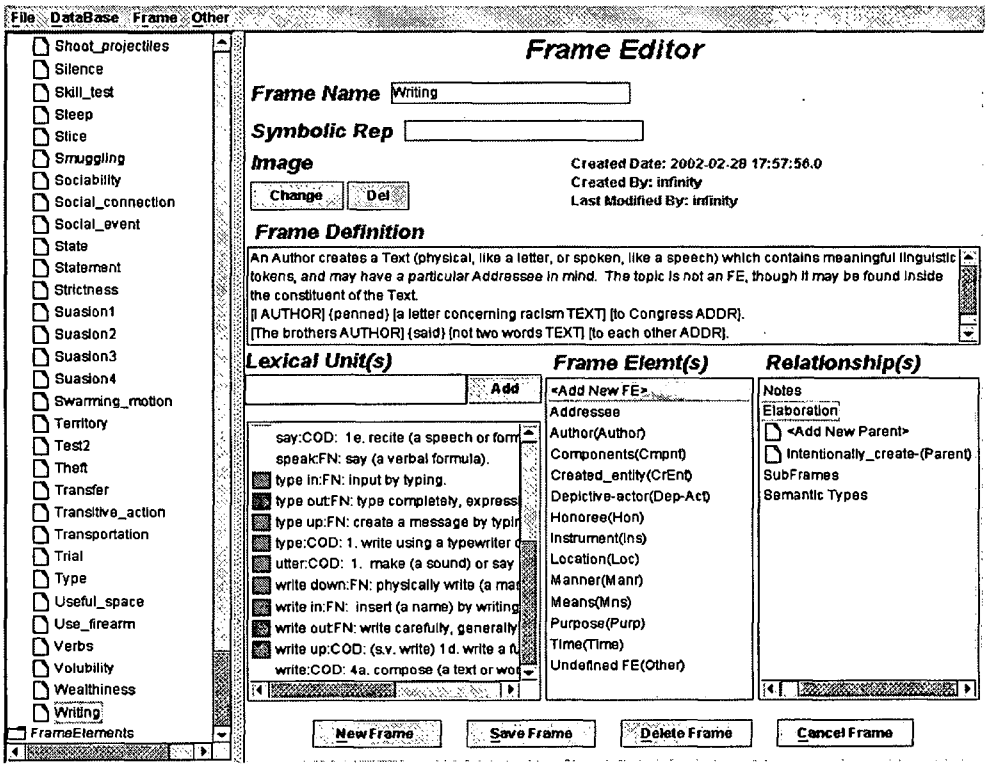

Figure 3. Screenshot of the Frame Editing Tool

The main view of the Frame Editor is shown in Figure 3. The editor is divided into a left and right frame, the former containing a listing of all available frames, the latter containing the information for the currently selected frame. Frame Definitions, including examples, are entered in the text box at the top. The lexical units belonging to the active frame are listed in the lower left box of the right frame below the field where new lexical units are entered. The Frame Elements defined for the frame are listed in the middle text box and relationships of the current frame to other frames can be indicated in the relationships text box. Adding new lexical units, frame elements, and new relationships calls up separate editors. Of these, the editor for indicating inheritance relationships is shown in Figure 4.

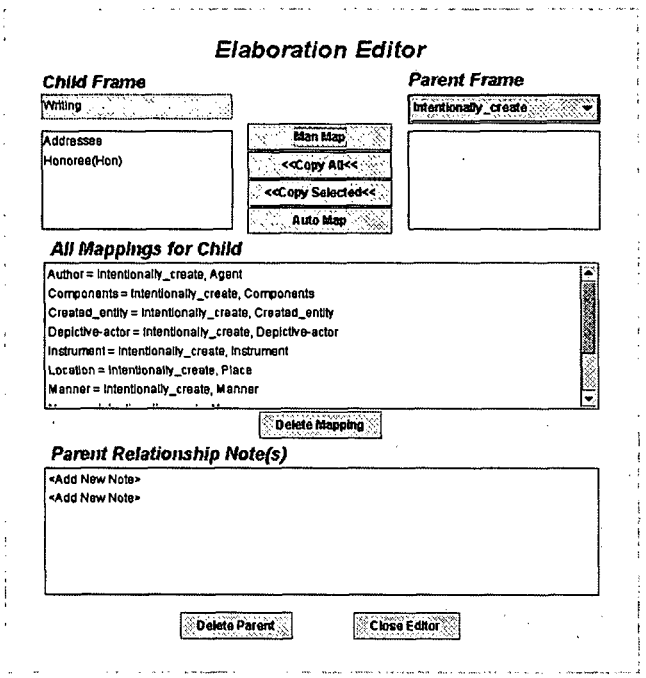

Figure 4. Screenshot of the Frame Inheritance Editing Tool

The relationships editor allows the lexicon analyst to record inheritance relationships between the current frame and other frames. As figure 4 shows, one can consider the semantic scenario of writing as a sub-type of intentionally creating something. The (not necessarily complete) equivalences between the semantic roles of the two frames are indicated as so-called mappings: the author of the Writing frame corresponds to the agent of the intentional creation frame; etc.

Collocations and MWEs in a Sample Text

In order to get some idea of the MWEs present in typical texts, let us consider the following example of journalistic prose, taken from CNN.com/LAW, dated 14 February 2001.

> Washington (CNN)--- Alleged White House gunman Robert Pickett was arraigned Wednesday at a federal court in Washington and ordered held without bond. A federal magistrate informed Pickett of the charges against him---assaulting a federal officer with a deadly weapon, which carries a maximum of ten years in prison. The magistrate set a preliminary hearing for next Tuesday and ordered Pickett held without bond. Pickett, who was shot in the knee by the Secret Service after allegedly firing two shots outside the White House, used crutches to walk into the court. He did not enter a plea.

A general way of applying FN valence information to the analysis of a sentence is to (1)

choose a word (starting from the highest semantically-relevant predicate), (2) determine the frames that this word evokes in this context, (3) notice the semantic roles of the participants in each such frame, (4) match the semantic needs associated with each such frame (hence with each sense of the word) with phrases found in the sentence, (5) those which permit the most coherent fit, and (6) register the semantic structures associated with the dependent constituents as provided by the selected frame.

But the analysis cannot simply proceed on the basis of frame information built on the words of the text taken one at a time. Many word sequences in our text must be identified as fixed phrases or tight collocations, the most obvious ones being the proper names *White House*, *Robert Pickett*, and *Secret Service*, others including *held without bond*, *assaulting a federal officer with a deadly weapon*, *preliminary hearing*, *firing shots*, and *enter a plea*. All of these phrases are parsable and semantically transparent, but they are also entrenched: *held without bond* is one of the standard phrases for reporting a decision in an arraignment hearing, *assaulting a federal officer with a deadly weapon* is a named offense in American law, *preliminary hearing* is a named step in the criminal justice process, and *fire* and *enter* are best treated as support verbs for the event nouns *shot* and *plea* respectively.

Many of these words evoke subframes of a frame involving steps in U.S. criminal process. Other phases of the process deal with bail, findings of guilt or innocence, sentencing, etc., and various abortions or alternative routings through the process such as skipping bail, changing one's plea, having a judge dismiss (or the prosecutor withdraw) the charges against the defendant, and so on. The lexical units in a complex frame will simultaneously evoke both the phase of the process which maps onto the grammatical structure of sentences containing a given lexical unit, and the larger event type of which that phase is a part.

Information on Multiword Expressions in the FrameNet Database

Information about Multiword Expressions is represented in, or derivable from, the FrameNet database in a variety of ways.

1. Multiword Lexical Units

Certain lemmas in the FN database were entered as multiword units from the start. Examples of MWEs entered as such will include lexicalized noun-noun compounds (wheel chair, etc.); verb-particle lemmas (trip up, etc.), and various kinds of idioms (cook someone's goose, etc.). In some cases lemmas originally treated on their own were later recognized as bested treated as part of a MWE and the analysis was changed accordingly.

2. Collocations Involving Subcategorization Details

In extracting sentences for annotation, lemmas are targeted in their subcategorizational contexts, and some of these involve the marking of a constituent by particular. Thus, we searched for the verb *object* and the noun *objection* in contexts where it preceded a to-phrase (nobody objected to your decision; the main objection to that proposal). These LUs can occur alone, where the entity objected to is missing but pragmatically salient (*I object! How could there be any objection*?). But whenever this FE is made explicit in the sentence, the syntactic constituent that expresses it has to be a prepositional phrase headed by *to*. (In this regard, then, verbs like *object* differ from the particle verbs, whose particles can never be

omitted.) To prepare for the discovery of such collocations (as *object to, prevent from, interested in, fond of*, etc.), the lexicographers identify the **core FEs** in the words' valence. That is, for each LU, we identify those FEs which are most centrally connected with the word's meaning, as distinguished from those that are more peripheral. Thus, in a sentence like *She objected to the bill in an angry editorial in the Times*, the phrase *to the bill* is more central than *in an angry editorial in the Times*. We can then regard LUs with lexically specified prepositions in their core valence as instances of multiword units.

3. Noun Compounds with Core FE as Modifier

Among the possibilities for the amplification of the frames associated with a noun is modification by another noun (*navy captain*) or by a pertinative[2] (also called relational) adjective (*naval commander*). Extracting noun compounds from the FrameNet database, then, will seek out either noun sequences identified as multiword targets, or nouns with modifiers that are labeled with core FEs.

4. Collocations across Transparent Nouns

One of the families of noun types we have tagged, which we call **transparent nouns**, includes nouns designating types, aggregates, parts, portions, classifiers, unitizers, etc., especially as they occur as the first noun in an *N of N* construction. By "transparent" we mean that in this construction the first noun is transparent with regard to collocational or selection relations between the second N and the external context of the construction [Fontenelle 20xx] or transparent to number agreement [Svensson 1998].[3] Examples with the relevant collocations underlined follow:

(1) In the 1920s, after the British literary establishment had neglected him for forty years, Machen attracted a **coterie** of admirers in the United States.
(2) Certain **strains** of Escherichia coli (E. coli), for example are responsible for causing "Traveller's Diarrhoea".
(3) He has pinned a little **square** of material onto both his knees so that when he drives, the fabric of his best trousers will not rub against the steering wheel.

5. Collocations with Transparent Nouns

The fact that we have transparent nouns labeled as such makes it possible to produce tables of these N-N pairs, and when we do we will find some that are lexically significant: *flock of geese, pride of lions, swarm of bees, bout of the flu, case of hepatitis*, etc. Of the Type nouns we will find many that are completely general (*type, kind, sort*), others that are more special (*variety, brand, strain, breed*). Thus, the decision to give special status to transparent nouns contributes to the detection of MWEs in two ways: first, there can be lexically relevant pairings between the two nouns in the construction, but secondly, a means of collecting linguistically relevant collocates can be devised in which it can be shown that the second N, not the first, figures in the collocational relation. (Thus, using examples in section 5, we can detect the collocations *attract admirers, pin material onto his knees*, and *Escherichia coli (E.coli) causes Traveller's Diarrhoea.*) This is related to the kernel dependency graph extraction exercise discussed below.

6. Event Nouns with Support Verbs

Among the contexts in which FrameNet annotators identify external arguments for event nouns is that of being a syntactic argument of an accompanying predicate (control structures, but also support verbs of all types). Since support verbs [Akimoto 1989; Mel'cuk 1995, 1996, 1998] represent the interesting case of objects selecting verbs rather than verbs selecting objects, we will find that there are lexicographically interesting pairings of support verb plus noun, and it will be possible to construct such tables as the following for the nouns in our Statement frame, which show these relationships:

Support Verbs	**Event Nouns**
make	address, admission, allegation, announcement, assertion, comment, **complaint**, concession, confession, declaration, exclamation, proclamation, remark, statement
give	address, exclamation, lecture
deliver	address, lecture
issue	declaration, denial, proclamation
utter	exclamation, remark
express, lodge, register, submit, voice	**complaint**
face, get	**complaint**
have	**complaint**, revelation

Table. 1 Event Nouns and Associated Support Verbs In the Statement Frame

Table 1 provides several interesting pieces of information. It shows that *make* occurs with the broadest range of nouns in this frame. It also suggests that the type of speech events that are *delivered* are ones that have a public audience rather than just an interlocutor as an Addressee. In addition, the table shows, in an indirect way, that various semantic roles of an event noun can be realized as the subject of different support verbs. Consider the fact that the noun *complaint* occurs in four different rows in the table. The verbs in the first two of the rows take the perspective of the speaker (*make; express, utter, lodge, register, submit*), whereas those in the last two rows (*face, get; have*) take the perspective of the addressee. This distinction is exemplified for *complaint* by the sentences in (4-5) and (6-7).

(6) The woman MADE no **complaint** to the police for five days.
(7) Voters have VOICED **complaints** at the elections being held before the trials begin, and before Mr Papandreou has a chance to prove his innocence.
(4) Bernard Antoine, general manager of the Novotel, West London, said he had RECEIVED no **complaints** about charges.
(5) Some of the things he says are really quite outrageous; do you ever GET any **complaints** about his language?

Inspection of support verb patterns across many semantic frames would likely support a larger (currently only intuition-based) generalization that *face* and *get* always express patients (or at least, non-agents).

Kernel Dependency Graphs (KDGs)

One of the side activities of the FrameNet work is that of devising a means of extracting what we are calling **kernel dependency graphs**, by which we mean displays of frame-bearing lexical units found in the corpus together with the lexical heads of the constituents that realize their core FEs. (in the case of phrases marked with function words, we would want this to include information about the marker and the head of the marked constituent. This would effectively be a display of governors together with their dependents along with an indication of both the semantic roles and the marking of those dependents. Thus, for a sentence like (8) the top-level KDG could be represented as in (9).

(8) The patient objected strenuously to the diet her doctor put her on.

(9) *object*
actor: patient
content: diet [**marker**: to]

One of the advantages of recognizing that arguments can be found at a distance from the predicates they are semantically dependent on, through control structures of the familiar kind, support verbs and transparent nouns, is that it becomes possible to zero in on the semantically correct KDGs in the data. Thus, from a sentence like (10) it should be possible to detect a KDG as in (11), centered in the noun *objection* that looks almost exactly like the previous one, by 'seeing through' the control structure around *likely*, the support verb *have*, and the transparent nouns; given the meaning of the sentence, *patient* and *diet* are more appropriate lexical companions to *objection* than *kind* and *sort*.

(10) That kind of patient is likely to have strong objections to this sort of diet.

(11) *objection* [**support**: have]
actor: patient
content: diet [**marker**: to]

The minimal parsing needed for finding the head nouns can generally be done automatically. By ignoring all of the transparent structures, we can easily find the words needed for extracting the semantically significant KDGs in a text. Of particular interest to our present point, some of the KDGs we are now able to recognize will turn out to be important collocations in their own right. Since special collocations occur not only between verbs and deverbal nouns and their complements, but also between adjectives and the nouns they modify, expanding the search for KDGs beyond complementation structures to modification structures allows us to add a new class of collocations. Including examples from Fontenelle [1999, pp. 28-29] we can find, by skipping past the transparent nouns, the collocations in the left column in the phrases given in the right column of Table 2. The first two are relations between verbs and objects; the next two are relations between adjectives and their semantic heads; the last two are relations between prepositions and their object nouns.

Collocation	Text
lay eggs	The hens laid dozens of eggs.
suffered fever	suffered a bout of fever
sound advice	a sound piece of advice
a fine mess	a fine sort of mess
on table	on this part of the table
in closet	in this part of the closet

Table 2. Finding Collocations across Transparent Nouns.

Summary

Some information about multiword expressions is a part of the FrameNet database because lexicographers chose to enter space-separated words as lexical units; the rest is derivable by searches or reports based on the information entered into the database by other means. Sometimes an argument (typically a subject) of a verb that takes a frame-bearing noun as its direct object is necessarily identified with a frame element of that object noun. While many such verbs (support verbs and other sorts of lexical functions in the sense of Mel'cuk) add configurational information of one kind or another to the verbal concept (features of aspect, point of view, evaluation, etc.), their main function in many cases is to combine with the nominal object to express a verbal meaning: all such pairings (verb + object noun) can count as MWEs; those in which the verb is lexically selected by the noun are entrenched MWEs and should be listed separately in the lexicon. Verbs, adjectives and nouns whose semantically basic complements are lexically specified as being marked by particular prepositions, can be counted as MWEs and listed with their prepositions. All noun compounds are MWEs, but only those which have meanings assigned to them beyond whatever semantic structure they may have by compositional principles will need separate entries in the lexicon. In sum, the database resulting from the straightforward lexicographic practice created for the FrameNet project has proved capable of yielding reliable information about participation in collocational patterns and multiword expressions for the words covered in the database.

Acknowledgements

We are grateful to the National Science Foundation for funding the work of the FrameNet project through two grants, IRI 9618838 "Tools for Lexicon Building" March 1997--February 2000, and ITR/HCI 0086132 "FrameNet++: An On-Line Lexical Semantic Resource and its Application to Speech and Language Technology" September 2000--August 2003. The Principal Investigators of FrameNet++ are Charles J. Fillmore, (ICSI), Dan Jurafsky (University of Colorado at Boulder), Srini Narayanan (SRI International/ICSI), and Mark Gawron (San Diego State University).

References

[Akimoto 1989] Akimoto, Minoji 1989, *A Study of Verbo-Nominal Structures in English.* Shinozaki Shorin, Tokyo.

[Baker et al. 1998] Baker, C.F., C.J. Fillmore & J.B. Lowe, 1998. The Berkeley FrameNet Project, in: *COLING-ACL İ98: Proceedings of the Conference, held at the University of Montreal*, pp. 86-90, Association for Computational Linguistics, Montreal.

[Fillmore 1985] Fillmore, C.J. 1985, Frames and the Semantics of Understanding, in: *Quaderni di Semantica* VI.2

[Fillmore & Atkins 1998] Fillmore, C.J. & B.T.S. Atkins 1998, FrameNet and Lexicographic Relevance, in: *Proceedings of the First International Conference on Language Resources And Evaluation.* Granada, Spain.

[Fillmore & Baker 2001] Fillmore, C.J. & C.F. Baker 2001, Frame Semantics for Text Understanding, in Proceedings of WordNet and Other Lexical Resources Workshop, held at North American Association for Computational Linguistics, Pittsburgh.

[Fillmore et al. 2001] Fillmore, C.J., C. Wooters & C.F. Baker 2001, Building a Large Lexical Database Which Provides Deep Semantics, in: B. Tsou & O. Kwong (eds.), *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation.* Hong Kong.

[Fontenelle 1999] Fontenelle, T. 1999, Semantic Resources for Word Sense Disambiguation: A Sine Qua Non?, in: *Linguistica e Filologia* 9. University degli Studi di Bergamo, Italy.

[Gildea & Jurafsky 2000] Gildea, Daniel and Daniel Jurafsky. 2000, Automatic Labeling of Semantic Roles, In Proceedings of the ACL 2000, Hong Kong.

[Mel'cuk 1995] Mel'cuk, I. 1995, Lexical Functions, in: L. Wanner (ed.) *Lexical Function in Lexicography and NLP.* Benjamins,

[Mel'cuk 1996] Mel'cuk, I. 1996, Phrasemes and Phraseology, in: M. Everaert, E.-J. van der Linden, A. Schenk, R. Schreuder (eds.), *Idioms. Structural and Psychological Perspectives*, Lawrence Erlbaum Associates. New Jersey.

[Mel'cuk 1998] Mel'cuk, I. 1998, Collocations and Lexical Functions, in: A. Cowie (ed.) *Phraseology. Theory, Analysis and Applications.* Clarendon Press.

[Svensson 1998] Svensson, P., 1998. *Number and Countability in English Nouns.* Uppsala: Swedish Science Press.

Endnotes

[1] The word "frame" here means "section of a window on the screen"!

[2] Pertinative adjectives ("pertainyms" in WordNet terminology) generally are not used predicatively and when modifying nouns generally function in ways similar to modifying nouns in noun compounds. (Compare *linguistic [adj] society* and *linguistics [n] society*, *Paris [n] connection* and *French [adj] connection.*) Occasions of predicate use are found in special constructions (*this problem is economic in nature*).

[3] Of course we need to recognize that not every instance of an *N of N* pattern is a transparent noun structure: many relational nouns occurring as the first N in this pattern can have a following *of*-phrase as a complement. Where the same word occurs in either such structure we can have local ambiguity, with cases in which it is the first noun of an *N of N* phrase that is the relevant collocate of something in its environment: compare *eat a number of apples* with *calculate the number of apples*. The noun *number* is a paradigmatic fellow to such nouns as *bunch*, *group*, *cluster*, *collection*, etc., in the one context, and to *quantity*, *size*, *weight*, *height*, etc., in the other.